Basic statistic concepts review – Exercises

(required files: almanac.csv (6), bank.csv (7) and Olympics Seoul.csv (8))

Each time that the computation of a correlation coefficient (exercises 1-3) is required do it:

- By “hand”, using a calculator;
- Using Excel;
- Using R

1. The Spearman's Rank Correlation Coefficient is used to discover the strength of a link between two sets of data. This example looks at the strength of the link between the price of a convenience item (a 50cl bottle of water) and distance from the Contemporary Art Museum(CAM) in El Raval, Barcelona.

Store	1	2	3	4	5	6	7	8	9	10
Dist (m)	50	175	270	375	425	580	710	790	890	980
Price	1.80	1.20	2.00	1.00	1.00	1.20	0.80	0.60	1.00	0.85

a. What can you conclude?
b. Can Pearson’s correlation coefficient be used with the same purpose? If yes do it, if no explain why.

2. Ten seventh-grade children randomly selected from a certain public school system were ranked according to the quality of their home environment and the quality of their performance in school, both rankings using a qualitative scale form 1 (worst) to 10 (best). The idea is to investigate a possible link between these 2 variables.

Child	1	2	3	4	5	6	7	8	9	10
Home	3	7	10	9	2	1	6	4	8	5
School	1	9	8	10	3	4	5	2	6	7

a. Calculate Spearman’s Rank correlation coefficient
b. Should Pearson’s correlation coefficient be used? If yes do it, if no explain why.

3. (Newbold, Carlson and Thorne) The accompanying table shows, for a random sample of 20 long-term growth mutual funds, percentage return over a period of 12 months and total assets (in millions of dollars)

Fund	1	2	3	4	5	6	7	8	9	10
Return	29.3	27.6	23.7	22.3	22	19.6	17.6	16.0	15.5	15.2
Assets	300	70	3004	161	827	295	29	421	99	756

Fund	11	12	13	14	15	16	17	18	19	20
Return	15.0	14.4	14.0	13.7	12.9	11.3	9.9	7.9	6.7	3.3
Assets	730	436	143	117	75	610	264	27	71	719

a. Calculate Spearman Rank correlation coefficient (using R, using Excel
b. Calculate Pearson correlation coefficient
c. Discuss the advantages (and disadvantages) of each solution
d. Is the computations done enough to conclude about a relation in the population?

4. Given $\overline{x}=\begin{bmatrix}12.45\\1.35\end{bmatrix}$ and the observed covariance matrix $\hat{S}=\begin{bmatrix}65.41 & 4.57\\4.57 & 1.27\end{bmatrix}$ answer the following questions

a. Determine the sample principal components and their variances.
b. Compute the loadings of the variables.
c. What interpretation, if any, can you give to the first principal component? (Assume that x_1 is the return on income and x_2 the earnings before interest and taxes)
d. Would the results change if correlation matrix is used to extract the principal components? Why? (answer this question without computing the principal components)

5. Consumers intending to purchase an automobile were asked to rate the following benefits desired by them in an automobile:

1) My car should have sleek, sporty looks.
2) My car should have dual air bags.
3) My car should be capable of accelerating to high speeds without seconds.
4) My car should have luxurious upholstery.
5) I want excellent dealer service.
6) I want automatic transmission in my car.
7) I want my car to have high gas mileage.
8) I want power windows and power door locks in my car.
9) My car should be the fastest model in the market.
10) I want to impress my friends with the looks of my car.
11) My car should have air conditioning.
12) My car should have AM/FM radio and cassette player installed.
13) I want my car dealer to be located close to where I live.
14) I want tires that ensure safe driving under bad roads conditions.
15) My car should have power brakes.
16) The exterior color of my car should be compatible with the upholstery color.
17) My car should have a powerful engine that provides fast acceleration.
18) My car should be equipped with safety belts.

19) My car should come with a service warranty that covers all the major parts.

Respondents indicated their agreement with the above statements using a five-point scale (1=strongly disagree to 5=strongly agree). The following table gives the loadings of the benefits on the principal components with eigenvalues greater than one.

Benefits	Prin 1	Prin 2	Prin 3	Prin 4	Prin 5
1	0.753	0.211	0.125	0.231	0.126
2	0.252	0.152	0.702	0.001	0.014
3	0.014	0.762	0.114	0.025	0.056
4	0.310	0.411	0.014	0.683	0.008
5	0.215	0.012	0.005	0.114	0.902
6	0.004	0.003	0.215	0.723	0.104
7	0.515	0.187	0.210	0.056	0.102
8	0.285	0.241	0.298	0.853	0.201
9	0.312	0.825	0.331	0.152	0.005
10	0.851	0.216	0.025	0.004	0.310
11	0.141	0.265	0.001	0.675	0.008
12	0.120	0.305	0.002	0.069	0.025
13	0.015	0.411	0.214	0.145	0.699
14	0.341	0.012	0.896	0.214	0.014
15	0.421	0.001	0.222	0.598	0.104
16	0.672	0.056	0.017	0.009	0.025
17	0.122	0.803	0.105	0.056	0.017
18	0.301	0.219	0.692	0.012	0.112
19	0.111	0.212	0.210	0.178	0.707

From the loadings given above identify the benefits that contribute significantly to each principal component and label the principal components. What therefore are the key dimensions that are considered by prospective car buyers?

6. In the Places Rated Almanac (file almanac.csv) Boyer and Savageau rated 329 communities according to the following nine criteria:

1) Climate and Terrain
2) Housing
3) Health Care and the environment
4) Crime
5) Transportation
6) Education
7) The arts
8) Recreation
9) Economics

For all but two of the above criteria, the higher the score the better. For housing and crime, the lower the score the better. The scores are computing using the following statistics for each criterion:

- **Climate & Terrain**: very hot and very cold months, seasonal temperature variation, heating- and cooling-degree days, freezing days, zero-degree days, ninety-degree days.
- **Housing**: utility bills, property taxes, mortgage payments.
- **Health Care & Environment**: per capita physicians, teaching hospitals, medical schools, cardiac rehabilitation centers, comprehensive cancer treatment centers, hospices, insurance/hospitalization costs index, fluoridation of drinking water, air pollution.
- **Crime**: violent crime rate, property crime rate.
- **Transportation**: daily commute, public transportation, Interstate highways, air service, passenger rail service.
- **Education**: pupil/teacher ratio in the public K-12 system, effort index in K-12, accademic options in higher education.
- **The Arts**: museums, fine arts and public radio stations, public television stations, universities offering a degree or degrees in the arts, symphony orchestras, theatres, opera companies, dance companies, public libraries.
- **Recreation**: good restaurants, public golf courses, certified lanes for tenpin bowling, movie theatres, zoos, aquariums, family theme parks, sanctioned automobile race tracks, pari-mutuel betting attractions, major- and minor-league professional sports teams, NCAA Division I football and basketball teams, miles of ocean or Great Lakes coastline, inland water, national forests, national parks, or national wildlife refuges, Consolidated Metropolitan Statistical Area access.
- **Economics:** average household income adjusted for taxes and living costs, income growth, job growth.

a. Carry out a PCA using R, assessing how many components should be considered in the analysis. If possible, label each retained component. Identify unusual cities.

b. As the 9 criteria are strongly skewed to the right, a colleague told you to transform your data before doing any analysis. So apply a log transformation (decimal logarithms) to the data set and then repeat the analysis developed un the previous question. Compare the results.

7. File bank.csv gives the correlation matrix of data from a customer satisfaction survey undertaken by ABC Savings Bank for their EasyBuy credit card. 540 respondents indicated their level of agreement/disagreement and level of satisfaction/dissatisfaction to the following 15 statements/services (Note that the correlation matrix is based on fictitious data).

Statements for The Customer Satisfaction Study for ABC Savings Bank

1. The bank processed my application for EasyBuy very quickly.
2. I found the facility of applying for EasyBuy over the phone extremely convenient.
3. Interest rates on EasyBuy are lower than most other credit cards.

4. I get my billing statements promptly at the beginning of each month.
5. EasyBuy has a very reasonably priced ``Wallet Protection Plan".
6. When I applied for EasyBuy, my credit rating approval was handled much faster compared to when I applied for my other credit cards.
7. EasyBuy customer service representatives are very knowledgeable.
8. The monthly billing statements are always accurate.
9. There is relatively less fluctuation in the interest rates of EasyBuy compared to most other credit cards.
10. Customer service representatives of EasyBuy are courteous and helpful.
11. I am allowed enough time to mail my payment.
12. The billing statements are detailed and easy to understand.
13. The money saving coupons accompanying the monthly billing statements offer great deals.
14. Its easy to reach the EasyBuy customer service representatives.
15. The bank allows interest rate reductions if total charge amount exceeds 1000 a year.

a. Perform a principal components analysis and identify the smallest number of components that best account for the variance in the data.
b. Label, if possible, the retained components.

8. Dataset Olympics (file olympics Seoul.csv) consists of the results from the 1988 Olympic women's heptathlon competition in Seoul. The variables are the following:

1) hurdles - results 100m hurdless
2) highjump - results high jump
3) shot - results shot
4) run200m - results 200m race
5) longjump - results long jump
6) javelin - results javelin
7) run800m - results 800m race

Using R, answer to the following questions

a. Using bivariate plots and correlation coefficients analyze a possible relationship between pairs of events. Is the sign of the correlation coefficients as expected?
b. To carry on a PCA should we scale (or not) the data set? Explain. Answer to the next questions according to your answer.
c. Now, run a PCA and make the scores plot
    i. How many principal components should you use?
    ii. Interpret the meaning of the retained components? What is the relation to the correlation matrix?
    iii. Can any athlete be considered as outlier?